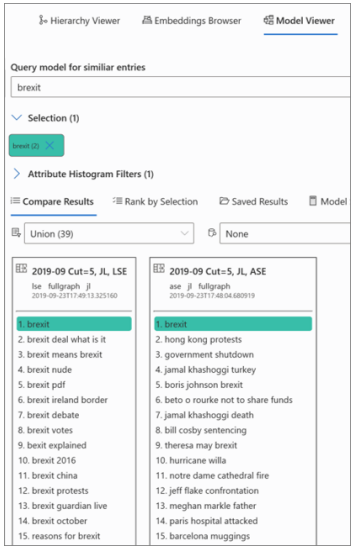

Figure 1: Model Viewer – one of three interfaces we have developed to view and compare embeddings of navigation graphs induced from search engine query transitions. While the LSE (Laplacian spectral embedding) model to the left shows different sub-intents for the query “brexit” (e.g., “brexit deal what is it”, the ASE (Adjacency spectral embedding) model to the right shows associated events of comparable public interest (e.g., “government shutdown” and “hong kong protests”). These different kinds of similar query may be used to drive different kinds of query recommendation in the search engine user experience.

Making Sense of Search: Using Graph Embedding and Visualization to Transform Query Understanding

Jonathan Larson
Microsoft Research
Silverdale, WA, USA
jolarso@microsoft.com

Nathan Evans
Microsoft Research
Silverdale, WA, USA
naevans@microsoft.com

Darren Edge
Microsoft Research
Cambridge, UK
daedge@microsoft.com

Christopher White
Microsoft Research
Redmond, WA, USA
chwh@microsoft.com

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04
http://dx.doi.org/10.1145//3334480.3375233

Abstract
We present a suite of interfaces for the visual exploration and evaluation of graph embeddings – machine learning models that reveal implicit relationships not directly observed in the input graph. Our focus is on the embedding of navigation graphs induced from search engine query logs, and how visualization of similar queries across different embeddings, combined with the interactive tuning of results through multi-attribute ranking and post-filtering (e.g., using raw query frequency or derived entity type), can provide a universal foundation for query recommendation. We describe the process of technology transfer from our applied research team to the Microsoft Bing product team, examining the critical role that visualization played in their decisions to ship the technology on bing.com.

Author Keywords
query logs; navigation graphs; graph embedding; query recommendation; visualization; tech transfer

CCS Concepts
•**Human-centered computing** → *Activity centered design; Visual analytics;* •**Information systems** → *Content ranking; Query log analysis; Query suggestion; Evaluation of retrieval results;*

Stage	Description
Awareness	Knowing of innovation
Interest	Seeking more information
Evaluation	Deciding on initial use
Trial	Learning from experiences
Adoption	Deciding on continued use

Table 1: The diffusion of innovations (Rogers, 1962 [7]).

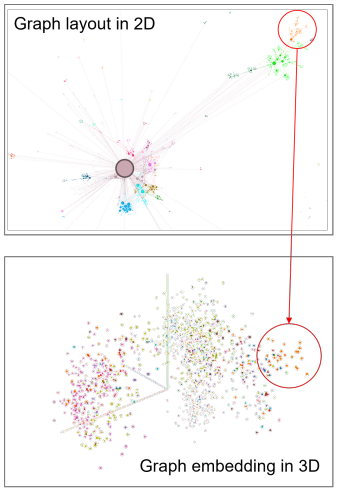

Figure 2: From graph layout in 2D (top) to graph embedding projected into 3D (bottom). The proximity of related vertices is preserved.

Introduction

While it is easy to make recommendations, it is hard to provide consistently good recommendations – especially when the questions can be vague and ill-formed, in any language, and on any topic. This is the fundamental problem of search engine query recommendation. A corresponding opportunity, arising from the growing dominance of mobile internet use, is that touch navigation of recommendation links is significantly faster and easier than manually typing new queries. Successive rounds of query recommendation also allow the user to define and refine their intent in an iterative, incremental, and exploratory fashion. Developing a universal solution for query recommendation thus has the potential to transform the search experience for all users.

This case study addresses the challenge of search query recommendation in the context of the Bing search engine, examining the critical role that data visualization and interactive data interfaces played in facilitating product group adoption of new recommendation mechanisms based on *graph embedding*. We describe the different qualities of the visual representations that drove the "diffusion" of this innovation (Table 1) from our applied research team in Microsoft Research, via key stakeholders in Bing product teams, to the broader Bing organization and onto bing.com.

The main lessons from this case study are twofold: (1) that the sensitivity of graph embedding to its configuration and the subjectivity of evaluating recommendation quality demands a shared medium for stakeholders to explore different options and discover best practices; and (2) that while *standalone explanatory data visualizations* may have sufficient power to develop awareness of and interest in new technologies, *shared exploratory data interfaces* may be necessary to drive the real-time evaluation by individual stakeholders that leads to collective trial and adoption.

Graph Embedding

Let $G = (V, E)$ represent a graph (network) of vertices (nodes) connected by edges (links), where the edges may optionally have a weight (strength) and direction (flow). Graph embedding [1] describes a family of machine learning techniques that take conventional *edge list* or *adjacency matrix* graph representations, which are hard to reason about on account of their discrete and high-dimensional structure, and transform them into *feature vector* vertex representations that are relatively easier to reason about because their continuous and low-dimensional structure defines a *metric space*. Graph embedding algorithms aim to perform this transformation in ways that preserve local structure, such that vertices sharing similar edges in the graph are mapped to similar locations in the embedding (Figure 2). The result is that rather than edges specifying the presence of a *relationship* between vertices, the distance between any pair of vertices can be interpreted as a measure of their *relatedness*. This enables simple spatial implementations of three fundamental inference tasks: 1) *vertex nomination* – given a query vertex, find its nearest neighbours in the embedded space; 2) *link prediction* – given a similarity threshold, find possible edges missing from the input graph; and 3) *community detection* – find clusters of vertices preferentially connected to one another.

While stochastic approaches to graph embedding use random walks to characterize vertex neighbourhoods before learning multidimensional representations (e.g., `DeepWalk` [5] and `node2vec` [4]), spectral approaches use eigendecomposition to factor matrix representations of the graph into a multidimensional set of orthogonal basis vectors. Different methods group vertices in different ways, e.g., Laplacian spectral embedding (LSE) groups vertices with similar connections, while Adjacency spectral embedding (ASE) groups vertices with a similar structural role (e.g., "hub") [6].

The transfer of graph embedding technologies as a diffusion of innovations

In the following series of sidebars, we provide a commentary on the process of technology transfer using quotes from two of our product group partners in Bing. We use the five stages of the "diffusion of innovations" to structure these observations.

***Awareness* – knowing of innovation**. Our partners knew of graph embedding and its possible use, but only in a general sense:

"There were teams doing similar things, just not using the technology that we're collaborating on and as a result, a lot of what they did was kind of handcrafted rather than a scaled solution." (P2)

"I do think that the ranking team were aware of graph embedding and may have been using it in specific cases, but not across the whole graph." (P2)

Navigation Graphs from Query Logs

As a way of understanding the relationships between search engine queries, we are interested in inducing *navigation graphs* that are represented implicitly in search engine query logs (Figure 3). These logs record both queries and clicks (on page results, query recommendations, and ads) for anonymous user sessions using Bing search. There is no one correct way to induce such a graph: the choice of vertex type (e.g., raw text vs normalized text vs named entities only), the relative weights given to different edge types (e.g., $query \rightarrow query$ vs $query \rightarrow recommendation$), and the minimum threshold for edge inclusion all make a significant impact on the character of the results.

The use of graph representations to understand co-occurrence relationships is a well established technique used in diverse domains ranging from literary analysis [10] to intelligence analysis [8]. The induction of navigation graphs from user history has also been practiced for at least two decades [2]. Such "memory-based recommender systems" use collaborative filtering in conjunction with association measures like pointwise mutual information (PMI) to rank observed transitions [9]. The promise of embedding-based approaches is that they are able to recommend relevant transitions even in areas where the transition signal is sparse. Recent work from Taobao [11] reports substantial sparsity in item transitions on the Taobao e-commerce platfrom (<1% of all possible transitions) and demonstrates that embeddings of such sparse navigation graphs can be used to generate recommendations whose click-through-rates (CTRs) significantly improve on the standard practice of collaborative filtering.

Data Interfaces for Making Sense of Search

Our first steps focused on the induction of navigation graphs from Bing query logs and the visualization of these graphs as node-link diagrams (Figure 3). These explanatory visualizations formed the basis of slides that we used to communicate the idea of navigation graphs and possible use cases for graph embedding. At the same time, we were developing a VS Code extension for visual exploration of graph embeddings in general (Figure 4). Our next step was to use this extension to explore Bing data in detail and to illustrate the mechanism of graph embedding to our partners.

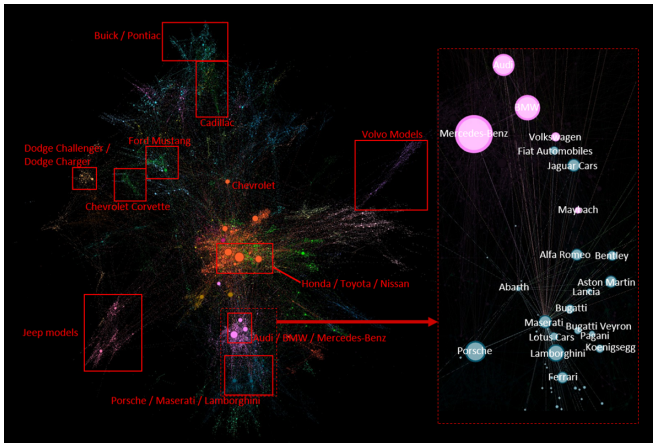

Figure 3: A node-link rendering of Bing query transitions in the automotive segment. As described by one of our partners: *"We just started with how people are transferring from query A to query B in the auto segment and then we are able to find luxury cars, or super cars, or we find that Jeep owners only care about Jeeps, or that people who shop for Corvettes also shop for Ford Mustangs – that's useful from a Bing experience as well as Ads perspective".*

The result of the multiple design iterations that followed was a cloud-based data platform and web-based data application for experimental evaluation of Bing navigation graphs and their embeddings. Using this platform, data scientists can deposit experimental research outputs (e.g., graph em-

***Interest* – seeking more information**. Our first graph visualizations persuaded our partners about the value of the approach, but the prevailing organizational culture was to persuade with data:

"I was just showing him how our users are navigating from entity to entity... He returned with this magnificent, wonderful, beautiful visualization of what the graph looks like and it was like, 'Wow, this feels like a much more useful and interesting dataset than we first thought'." (P1)

"For me, I was sold even at the very beginning with the initial tool [Figure 4], which was more of a strict visualization tool that shows you the whole of the graph color-coded... Right there, I got it, I totally understand what the value would be of having all of Bing in a graph." (P2)

"There was initially a bias where people would look at this and think it was [just] a visualization tool... but really people are not interested in visualization tools in and of themselves." (P2)

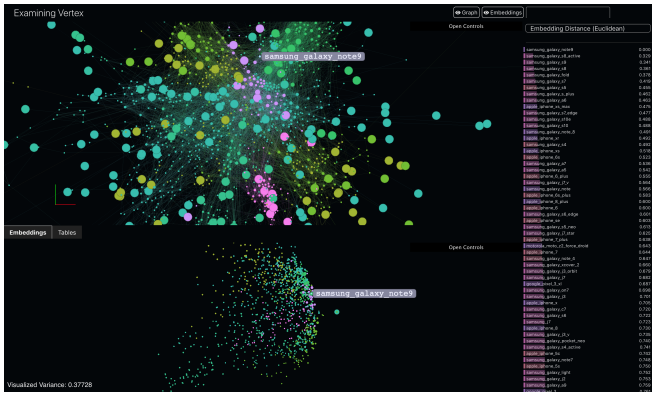

Figure 4: VS Code extension for exploring graph embeddings using coordinated views of a 2D node-link graph (top), a 3D embedding projection (bottom), and a ranked list of nodes (right).

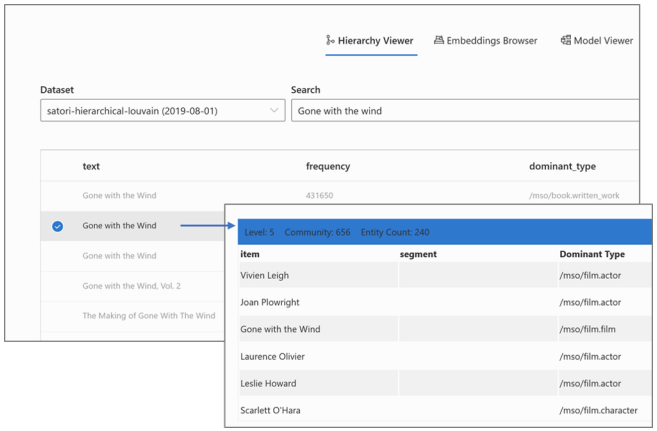

Figure 5: Hierarchy Viewer for the query "Gone with the wind", listing top entities in the matching leaf cluster. Includes a lead actor (Vivien Leigh), their character (Scarlett O'Hara), and their famous lover not cast in the film (Lawrence Olivier).

beddings and derived data such as hierarchical "communities" of related entities) such that they may be automatically indexed and joined with product group datasets (e.g., types and identifiers for knowledge graph entities) for interactive exploration in the browser. Our data application has three main interface tabs, introduced next.

With *Hierarchy Viewer* (Figure 5), the user can explore hierarchical communities of queries related to a given query, augmented with the query frequency and information on the dominant entity from our internal knowledge graph. Communities are detected from navigation graphs using a range of methods, including statistical (graph-based) and spatial (embedding-based) community detection techniques.

With *Embeddings Browser* (Figure 6), the user can explore many possible rankings of entity recommendations related to a given query, augmented with information from our knowledge graph and query logs (e.g., number of typed queries vs requery clicks). Our use of the `LineUp` visualization [3] allows the user to rerank the recommended entities based on any attribute (e.g., its own frequency vs its similarity to the query). The user can also create weighted combinations of attributes by dragging columns onto one another, before dynamically reweighting the attributes (and thus reranking the results) by directly manipulating relative column widths. Recommendations may also be filtered interactively (e.g., based on frequencies and/or entity types) to derive multiple kinds of recommendation, each with a distinctive character and purpose in terms user experience.

With *Model Viewer* (Figure 1), the user can load any subset of the embedding models for side-by-side comparison of the top N results across a series of queries. Result lists can be filtered to show their union (all items), intersection (common items), and symmetric difference (unique items).

***Evaluation* – deciding on initial use**. Experimentation persuaded segment owners to trial in production:

"When we meet with different segment owners, we show the tool and they really like playing with it, because they find for themselves and they understand for themselves that combining similarity with other signals works well in general... eventually they familiarize themselves with enough queries and results that they trust the data and give the go ahead to the engineers to take the raw data and try to scale this and flight this and ship this." (P1)

***Trial* – learning from experiences**. Personal evaluations outweighed prior results in other areas:

"the other segment owners, they all want to be individually convinced by more than just a flight... when I show them this view [Figure 6], it looks user friendly, it looks high quality, and they feel empowered to make their own decision." (P1)

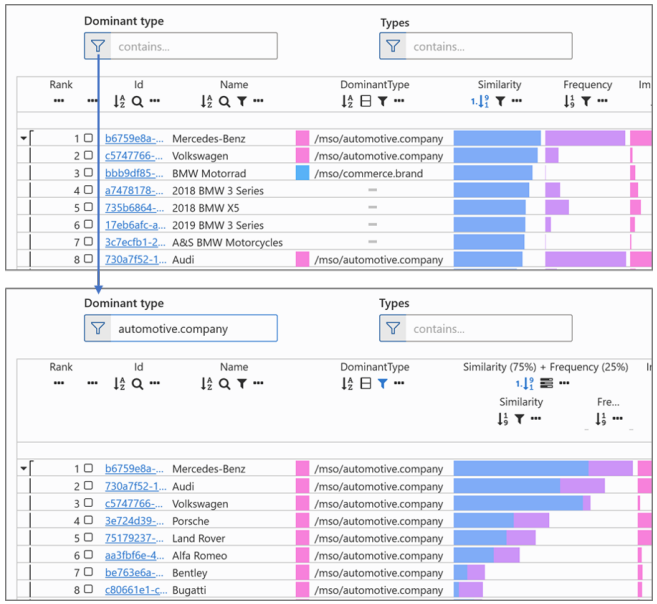

Figure 6: Embedding Browser for the query "BMW". Results ranked by similarity (top) are refined by jointly ranking on frequency (75%–25%) and filtering by entity type (bottom).

Toggling any item highlights that item in a common color across all juxtaposed models, while "Rank by Selection" ranks all models based on their coverage of the selected items (using the Jaccard set similarity measure of intersection over union). Finally, "Model Stats" presents a matrix comparing all loaded models to one another using either Jaccard similarity or average precision (treating the model represented by each row as the ground truth ranking for all other models). Together, these capabilities allow detailed comparison of many candidate models across a wide range of queries, allowing the user to understand which of the existing models works best for their needs, and which

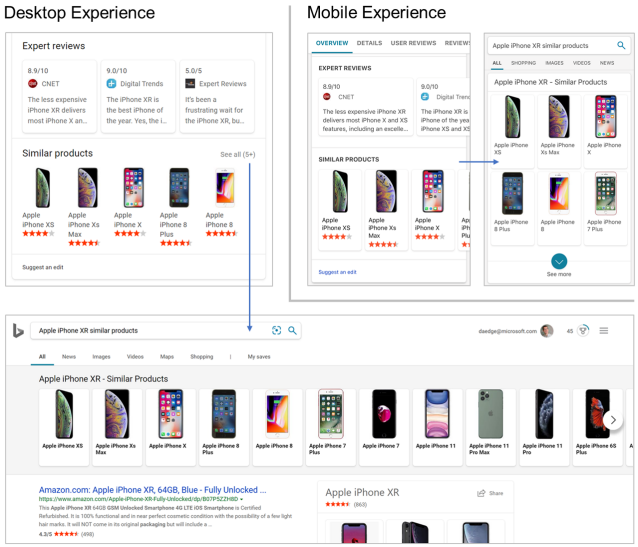

Figure 7: Similar products feature on bing.com, showing a carousel of recommended links for smartphones similar to the query "iPhone XR" (Country/Region: United States – English).

new models should be developed and evaluated next before flighting or releasing to production. Figure 7 shows the desktop and mobile experiences of one such feature shipped on bing.com – "Similar products" recommendation.

Finally, the web application offers a link to "Segment spreadsheets" – a SharePoint folder in which the top recommendations for all queries associated with each predefined search segment (e.g., health or automotive) are exported to their own Excel spreadsheets for custom analysis (e.g., by filtering and ranking in Excel, by importing to Business Intelligence platforms like Power BI and Tableau, or by performing data science using languages like R and Python).

***Adoption* – deciding on continued use**. Improvements in user experience and revenue ultimately secured continued investment:

"In the past we were just using PMI to do related entities, but the unique challenge in product segments is that there is a very sparse signal of users transitioning from one product to another. The cool thing now is the similarity score, where even if you don't have a lot of raw signal, we're able to find close relationships. So really it helped us to create the related entities feature that we couldn't have done just with raw data... and it's having a direct monetary impact on Bing, it's really exciting." (P1)

Overall, this analysis shows the power of exploratory data interfaces to overcome both the sensitivity of machine learning and the subjectivity of human decision makers, as well as the power of graph models and embeddings to reveal latent value in data.

How Visual Representations Drove Adoption

We interviewed two of our primary partners in Bing to understand their experience of the technology transfer process and how our interfaces have influenced sense-making and decision-making across the organization. Both are highly experienced program managers who have worked in a variety of Microsoft product groups before their current roles in Bing, where they are responsible for the development of new metrics and features, as well as the cross-group coordination with segment owners (e.g., for queries on health, sports, politics, food, etc.) required to ship new and improved features in production.

A central theme emerging from these interviews was that visual representations had played a vital role at each stage of the adoption process [7], which we have illustrated through a series of sidebars concluding in the left margin of this page. In the sections below, we reflect on how exploratory data interfaces helped to make sense of search.

Visual representations suggest new metrics and experiences
P1 described the close relationship between new visualizations and metrics inspired by newly-visible phenomena: *"The question is, 'What should the next metric be, and how should we get people to buy into that metric?' And really visualization is the key there, because if people can see the problem, they will want to act on it and track the improvement over time".* From *"the maybe 50-60 metrics"* he had created and presented to executives, navigation graph metrics had *"gained the most traction"*. One reason was the clear advantage that embedding-based methods had over the use of PMI, which was the prevailing standard practice: *"It's what you're blind to with PMI ... let's say you're on this specific node. PMI is just going to tell you, 'go here, go here, go here, go here... [to adjacent nodes]', but that's not the only thing the user should do. They should also move to nearby clusters".* This sense of 'bottom up' query clustering in ways that may be independent of 'top down' segment definitions, along with a sense for the spatial proximity of such clusters, both emerged from exposure to our early node-link visualization of navigation graphs (Figures 3 & 4).

Visual clusters reveal gaps in existing knowledge
P2 explained how one of the most important capabilities offered by our graph embedding approach was the automatic generation of *"reasonably clustered sets where we can kind of narrow down and identify how dense the entity graph is within that particular cluster"*. Tool users can thus interactively answer the questions, *"Are there missing entities in that particular cluster? Are the entities too generic? Do we need more specific entities in [our knowledge graph]?"*. At the same time, both partners critiqued how the query-driven, list-based representation in this view showed only a narrow slice through the navigation graph, and thus lacked the perceptual and navigational affordances of the node-link visualizations we had initially used to communicate the ideas of navigation graphs and graph embeddings (Figures 3 & 4). While our design choices with the current tool were guided by the 'search and list' paradigm most familiar to our partner team, this reinforces the idea that new tools can lead to cultural change (which the tools must then follow).

P1 gave a detailed account of how a Bing-scale navigation graph combined with the ability to experiment in real-time enabled a new kind of collaborative decision-making: *"Let's say I'm on a call and they say, 'I have this problem...', I can go there in real-time and try their specific query, their specific segment, tweak the ranking, the filtering, and show it to them in real-time... being able to explore all this intelligence from the dataset in real-time is valuable because you are riding the wave of the meeting you're in... Instead of saying, 'let's schedule another meeting', or 'I'll reply after the meet-*

ing with some extra insights', you can extract it in real-time and have a discussion in real-time... there's a lot of value in the tool but what I like the most is that I have access to the full dataset and I can explore it in real-time".

Visual juxtaposition builds understanding of methods
P1 called out Model Viewer as ending up *"the most useful to us"* as it allowed direct model comparison at scale, emphasizing their similarities and differences as well as their suitability to different problems: *"We can compare different embedding algorithms side-by-side and this is how we really understood that one algorithm is really good for typos, one is really good for sub-intent, one is really good for related entities. That was really a breakthrough there, and once you understand which embedding algorithm is good for what, then you need to try it with different permutations of hyperparameters for tuning, and this is the other part where this tool was useful, because then you can try like 200 permutations and figure out which one's best".* Showing an example query for 'Bill Gates', P2 also called out the different qualities of result lists arising from the use of Adjacency vs Laplacian spectral embedding: *"In the ASE model, you can see that this is a really tight set of entities, especially at the top. This would be the 'find missing entities and find which entities are related to a specific entity' cut. Whereas the LSE model shows a lot of re-queries with the term Bill Gates, which is so powerful... For example, you could look at the popular re-query terms for all the people that are like Bill Gates, and then you could rank those to get the best set of re-query terms for that particular entity type. And by entity type, Bill Gates would be people.person, but if you look at this [ASE] list you can clearly see we could create a new type of people.person.billionaire".* These insights – that multiple embedding methods can be used together to infer new entity types as well as their sub-intents – were both consequences of visual model juxtaposition.

Conclusion

In this case study, we have shown the value of both inducing and embedding navigation graphs as a novel approach to "making sense of search". While the resulting data assets may be a necessary foundation for universal query recommendation, such data alone are insufficient for driving decisions by product owners to trial the data and associated algorithms in production. In our experience, such decisions greatly benefit from the visual representation and exploration of data, especially when the data are comprehensive, the queries are relevant to the viewer, and the results are tunable based on subjective notions of quality.

At the same time, results that look good to software engineers and program managers in principle may not perform well with end-users in practice, and the continued use and broader adoption of navigation graph embeddings within Bing will always depend on their measured impact on both user engagement and revenue. Our models are driving experimental flights of both desktop and mobile experiences, as well as shipped experiences for "Similar products" and "People also search for" in a limited but expanding set of segments and markets. We are observing statistically significant improvements on the general metrics of session success rate, time to success, and estimated revenue per click, as well as recommendation clicks and carousel actions. We hope that our work continues to transform the Bing user experience for the better and that others may benefit from the lessons of this case study.

Acknowledgements

We thank our team members in Microsoft Research, our product group partners in Bing, and Carey Priebe and his team at the Johns Hopkins University Applied Mathematics and Statistics Department.

REFERENCES

[1] Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637.

[2] Xiaobin Fu, Jay Budzik, and Kristian J. Hammond. 2000. Mining Navigation History for Recommendation. In *Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI '00)*. ACM, New York, NY, USA, 106–112. DOI: http://dx.doi.org/10.1145/325737.325796

[3] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2277–2286. https://github.com/lineupjs

[4] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864. https://github.com/aditya-grover/node2vec

[5] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710. https://github.com/phanein/deepwalk

[6] Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. 2019. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences* 116, 13 (2019), 5995–6000.

[7] Everett M. Rogers. 1962. *Diffusion of innovations* (1st ed.). Free Press of Glencoe, New York. (book).

[8] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

[9] Eva Suárez-García, Alfonso Landin, Daniel Valcarce, and Álvaro Barreiro. 2018. Term Association Measures for Memory-based Recommender Systems. In *Proceedings of the 5th Spanish Conference on Information Retrieval*. ACM, 6.

[10] Romain Vuillemot, Tanya Clement, Catherine Plaisant, and Amit Kumar. 2009. What's being said near "Martha"? Exploring name entities in literary text collections. In *2009 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 107–114.

[11] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 839–848.